

APPLICATION OF A SECOND-ORDER STOCHASTIC OPTIMIZATION ALGORITHM FOR FITTING STOCHASTIC EPIDEMIOLOGICAL MODELS

Atiye Alaeddini

Daniel J. Klein

Institute for Disease Modeling
3150 139th Ave SE, Building 4
Bellevue, WA, 98005, USA

ABSTRACT

Epidemiological models have tremendous potential to forecast disease burden and quantify the impact of interventions. Detailed models are increasingly popular, however these models tend to be stochastic and very costly to evaluate. Fortunately, readily available high-performance cloud computing now means that these models can be evaluated many times in parallel. Here, we briefly describe PSPO, an extension to Spall's second-order stochastic optimization algorithm, Simultaneous Perturbation Stochastic Approximation (SPSA), that takes full advantage of parallel computing environments. The main focus of this work is on the use of PSPO to maximize the pseudo-likelihood of a stochastic epidemiological model to data from a 1861 measles outbreak in Hagelloch, Germany. Results indicate that PSPO far outperforms gradient ascent and SPSA on this challenging likelihood maximization problem.

1 INTRODUCTION

In public health, it is critical to have a reasonable understanding of an epidemic disease in order to set pragmatic goals and design highly-impactful and cost-effective interventions. Mathematical models of these epidemiological processes can support decision making by forecasting disease spread in space and time, and by evaluating intervention outcomes many times in-silico before spending valuable resources implementing real-world programs. Recent efforts in computational epidemiology have focused on the design and application of detailed stochastic models that capture physical mechanisms through which disease propagates, along with the statistical fluctuations inherent in complex systems. These stochastic models are readily available, and recent work has focused on applying these models to malaria (Eckhoff et al. 2016, Eckhoff 2013, Marshall et al. 2016, Gerardin et al. 2016), HIV (Bershteyn et al. 2016, Eaton et al. 2015), polio (McCarthy et al. 2016, Grassly et al. 2006), and more.

A central challenge in working with these complex models lies in finding input parameters that result in model outputs closely matching observed data. The stochastic nature of these models means that each set of input parameters maps to a distribution of outcomes, from which each sample (model run) can take several hours to obtain. In the quick-to-evaluate deterministic model setting, almost any classical optimization algorithm, such as steepest descent or Newton-Raphson, can be used to fit the model to data by maximizing the pseudo-likelihood of the parameters given the data. However, care must be taken when applying deterministic methods to stochastic objective functions, as the inherent noise causes unexpected behavior. Some have tried model averaging, wherein the results at a given point are averaged over many replicates to reduce stochastic fluctuations. While this approach mitigates stochasticity, it is expensive and does not resolve the problem. Further, algorithms that evaluate the model one point at a time, like Markov chain Monte Carlo and most optimization algorithms, will simply take too long to converge when applied to these time- and memory-expensive models.

There exist many optimization methodologies for both deterministic and stochastic systems. The steepest descent method (Wright and Nocedal 1999) is the most prominent iterative method for optimizing a complex objective function. The gradient-based algorithms, such as Robbins-Monro (Robbins and Monro 1951), Newton-Raphson (Fröberg 1969), and neural network back-propagation (Rumelhart et al. 1985), rely on direct measurements of the gradient of the objective function with respect to the optimization parameter. But, in many cases the gradient of the loss function is not available. This is a common occurrence, for example, in complex systems, such as the optimization problem given in (Tsilifis et al. 2017), the exact functional relationship between the loss function value and the parameters is not known, and the loss function is evaluated by measurements on the system or by running simulation. A review on the main areas of optimization via simulation can be found in (Fu 1994, Hong and Nelson 2009, Swisher et al. 2000).

Some algorithms have been developed specifically to optimize stochastic black-box cost functions. Of these, the Kiefer-Wolfowitz algorithm (Kiefer et al. 1952) is perhaps the most well known. This algorithm estimates the gradient from noisy measurements using finite differencing. Another well known stochastic optimization algorithm in the case of high dimensional problems is Simultaneous Perturbation Stochastic Approximation (SPSA), which is an approximation algorithm based on simultaneous perturbation (Spall 1992). Later, J. C. Spall presented a second-order variant of SPSA (Spall 2000). The SPSA algorithm is used extensively in many different areas, e.g. signal timing for traffic control (Ma et al. 2013), and some large scale machine learning problems (Byrd et al. 2011). The convergence of this algorithm to the optimal value in the stochastic *almost sure* sense makes it suitable in many applications. Despite these advantageous properties, SPSA is a serial algorithm that evaluates the (stochastic) function a few points at a time. Specialized algorithms, like the PSPO algorithm (Alaeddini and Klein 2017) we employ here, are thus needed to take advantage of fundamentally-parallel resources like high-performance cloud-based computing. Other researchers have looked at ways of enhancing the convergence of the SPSA algorithm, e.g. iterate averaging is an approach aimed at achieving higher convergence rate in a stochastic setting (Polyak and Juditsky 1992). Another variant of SPSA in which the parameter perturbations are based on deterministic, instead of random, perturbations was considered in (Bhatnagar et al. 2003). Several other modified versions of SPSA were proposed in the literature (Kocsis and Szepesvári 2006, Spall 2009).

Among epidemiological parameter search algorithms, some have focused on finding *all* parameter combinations that are consistent with observed data. These algorithms, such as basic sampling with importance sampling, Bayesian History Matching (Andrianakis et al. 2015), and Incremental Mixture Importance Sampling (Wagner et al. 2014) have been used effectively. However, the challenging task of finding all parameter combinations means that these algorithms are only applicable in lower dimensional parameter spaces, typically 20 or less. Other approaches have focused on finding parameter combinations that achieve a binary goal-oriented outcome, e.g. disease eradication, at a specified probability (Roh and Eckhoff 2014, Klein et al. 2014). When the population counts are sufficiently large, mean field theory can be used to give an approximate model, popular among researchers who study disease spread in a network (Alaeddini and Morgansen 2016, Preciado et al. 2013, Miller et al. 2012). While all of these algorithms are useful in specific situations, none address the challenge of finding the best inputs, especially for models that are stochastic and have many parameters.

Uncertainty in stochastic modeling is either fluctuation-driven (intrinsic) or parameter-driven (extrinsic). Both sources of uncertainty are important to quantify, as the noise distributions can impact results (Bayati and Eckhoff 2012), and intervention impact can be sensitive to parameters. To approximate extrinsic uncertainty in an optimization framework, we make use of the Cramer-Rao Bound (Rao 1992). This bound is derived from the inverse of the Fisher Information Matrix, which evaluates the expected curvature of the log-likelihood at the point of maximum value. High curvature is indicative of easily identified parameters, and thus a tight error bound.

The main contribution of this paper is the application of PSPO, a novel second-order parallel-friendly algorithm, to fitting a stochastic model to historical measles data from Hagelloch, Germany. The PSPO algorithm we use here, take advantage of parallel cloud-based computing resources, which is appropriate

for the specific application described in this paper. We also use the relationship between the number of parallel rounds of computation and error tolerance of the gradient for each iteration in PSPO algorithm to achieve an acceptable convergence rate for this particular application. We demonstrate that PSPO works well for fitting a stochastic spidemiological model and compares favorably to the conventional simultaneous perturbation optimization algorithm (SPSA).

The rest of the paper is structured as follows. In Section 2, we give a summary of the simultaneous perturbation optimization method. The model under study is introduced in Section 3. A review of the modified stochastic optimization algorithm, called PSPO, is presented in Section 4. The numerical simulations are given in Section 5, and Section 6 concludes the paper.

2 PRELIMINARIES: SIMULTANEOUS PERTURBATION STOCHASTIC APPROXIMATION

Consider the problem of maximization of a reward function $L(\theta) : \mathbb{R}^p \rightarrow \mathbb{R}$. J. C. Spall proposed an efficient stochastic algorithm called Simultaneous Perturbation Stochastic Approximation (SPSA) (Spall 1992, Spall 2000, Spall 2009). The SPSA algorithm efficiently estimates the gradient and the Hessian matrix using finite difference techniques. This algorithm basically consists of two parallel recursions for estimating the optimization parameter, θ , and the Hessian matrix, $H(\theta)$. The first recursion is a stochastic equivalence of the Newton-Raphson algorithm, and the second one estimates the Hessian matrix. The two recursions are

$$\begin{aligned} \theta_{k+1} &= \theta_k + a_k (\Pi_{\mathcal{P}}(\bar{H}_k))^{-1} \hat{\mathbf{g}}_k(\theta_k), \\ \bar{H}_k &= \frac{k}{k+1} \bar{H}_{k-1} + \frac{1}{k+1} \hat{H}_k, \end{aligned} \tag{1}$$

where a_k is a positive scalar factor, \mathcal{P} denotes the set of all negative definite matrices, and $\Pi_{\mathcal{P}}(\cdot)$ is the projection into the admissible set \mathcal{P} . Here, $\hat{\mathbf{g}}_k$ and \hat{H}_k are the estimated gradient and Hessian at iteration k . The SPSA approach for estimating the $\hat{\mathbf{g}}_k(\theta)$ and $\hat{H}_k(\theta)$ follows.

Let $\Delta_k \in \mathbb{R}^p$ be vectors of p mutually independent zero-mean random variables satisfying the condition of $E\{\Delta_k^{-1}\}$ be bounded. An admissible distribution is a Bernoulli ± 1 distribution. The two-sided estimate of the gradient at iteration k is given by:

$$\hat{\mathbf{g}}_k(\theta_k) = \frac{L(\theta + c_k \Delta_k) - L(\theta - c_k \Delta_k)}{2c_k} \Delta_k^{-1}, \tag{2}$$

where c_k is a positive scalar. Note that in (2), Δ_k^{-1} is the element-wise inverse of Δ_k . For second order SPSA, it is suggested to using a one-sided gradient approximation, given by:

$$\hat{\mathbf{g}}_k(\theta_k) = \frac{L(\theta + c_k \Delta_k) - L(\theta)}{c_k} \Delta_k^{-1}. \tag{3}$$

Now let $\tilde{\Delta}_k \in \mathbb{R}^p$ be vectors of random variables satisfying the same condition of Δ_k . The positive scalars \tilde{c}_k are chosen to be smaller than c_k . The following formula gives a per-iteration estimate of the Hessian matrix.

$$\hat{H}_k = \frac{1}{2} \left[\frac{\delta G_k}{2\tilde{c}_k} \tilde{\Delta}_k^{-1} + \left(\frac{\delta G_k}{2\tilde{c}_k} \tilde{\Delta}_k^{-1} \right)^T \right], \tag{4}$$

where

$$\delta G_k = \hat{\mathbf{g}}_k(\theta_k + \tilde{c}_k \tilde{\Delta}_k) - \hat{\mathbf{g}}_k(\theta_k - \tilde{c}_k \tilde{\Delta}_k). \tag{5}$$

The practical implementation details of this algorithm are given in (Spall 2000, Spall 2012).

3 PROBLEM DEFINITION

In this paper, we formulate our problem for virus spread in a population consists of N individuals. In order to have a reliable predictive model, we need to estimate the model parameters using noisy prevalence data. Two main sources of prevalence data are clinics and health surveys. The clinic data are often available for several time periods. But, health survey data are usually available at only one or a few number of time points. We do not make any assumption on whether the prevalence data is sparse or dense. The output, ζ , is the prevalence rates at time points, t_1, t_2, \dots, t_n . To estimate the model parameters, we use maximum likelihood estimate (MLE). Assume the likelihood function of a given θ for a given observation ζ is expressed as:

$$l(\theta | \zeta_1, \zeta_2, \dots, \zeta_n) = \Pr(\zeta_1, \zeta_2, \dots, \zeta_n | \theta), \tag{6}$$

where, ζ_j is the observed prevalence at time t_j . Then the log likelihood function is given by

$$L(\theta) = \log l(\theta | \zeta_1, \zeta_2, \dots, \zeta_n) = \sum_{j=1}^n \log \Pr(\zeta_j | \theta). \tag{7}$$

In order to find the probability, we use a beta binomial model. The beta binomial model is a type of Bayesian model. In our model, the probability of having I infected individuals among N people can be expressed as a binomial distribution

$$I | N, \kappa \sim \text{Binomial}(\kappa, N)$$

where κ is the probability of each individual being infected. Thus the binomial likelihood is

$$\Pr(I | \kappa, N) = \frac{N!}{I!(N-I)!} \kappa^I (1-\kappa)^{N-I}.$$

Now, κ itself is a probability that can vary over the interval $[0, 1]$, and a Beta distribution is used as the prior

$$\kappa | \beta_1, \beta_2 \sim \text{Beta}(\beta_1, \beta_2),$$

then, after observing I infected individuals in a simulation, we have

$$\kappa | I, N, \beta_1, \beta_2 \sim \text{Beta}(\beta_1 + I, \beta_2 + N - I).$$

Using the beta binomial distribution, the probability of observing ζ_j given modeling parameter θ is

$$\Pr(\zeta_j | \theta) = \int_0^1 \Pr(\zeta_j | \kappa, N) \Pr(\kappa | I_j, N) d\kappa = \frac{N!}{\zeta_j!(N-\zeta_j)!} \frac{\Gamma(2+N)\Gamma(1+I_j+\zeta_j)\Gamma(1+2N-I_j-\zeta_j)}{\Gamma(1+I_j)\Gamma(1+N-I_j)\Gamma(2+2N)} \tag{8}$$

where, I_j comes from running our epidemic model on a population of N individuals for time $t \leq t_j$, and substituting θ as model parameters. In this paper, the evolution of an epidemic disease can be simulated using Gillespie's stochastic simulation algorithm (SSA) (Gillespie 2001). In fact, we employ a stochastic compartmental model, in which state transitions resemble chemical reactions, although the PSPO methodology works equally well on individual-based models.

Finally, the optimal model parameters, θ^* , are obtained by solving the following optimization problem:

$$\theta^* = \arg \max L(\theta),$$

where $L(\theta)$ is given by (7). In the next section we present a robust stochastic algorithm to estimate the optimal parameter, θ^* .

In this paper, we are also interested in calculating the confidence region of the estimated optimal solution obtained by our optimization algorithm. The Fisher information matrix is a tool used here to compute the confidence region. Given the information matrix, one can use Cramer-Rao inequality to compute the confidence region of the estimated parameter. Assuming θ^* is the true value of the parameter, then the uncertainty of the estimated parameter, $\hat{\theta}$, is bounded using the Cramer-Rao inequality as

$$\Sigma_{\theta} \geq F^{-1}(\theta^*),$$

where Σ_{θ} is the covariance matrix of $\hat{\theta}$, and F is the Fisher information matrix (Scharf 1991). In other words, the estimated standard errors of the maximum likelihood estimation are the square roots of the diagonal elements of the inverse of the observed Fisher information matrix. Basically, smaller information matrix means larger variation in the estimated parameter and less confidence in the accuracy of the estimated value. Having the likelihood function, the $p \times p$ Fisher information matrix is given by:

$$F(\theta) \equiv \mathbb{E} \left[\frac{\partial L(\theta)}{\partial \theta} \cdot \frac{\partial L(\theta)}{\partial \theta^T} \middle| \theta \right], \tag{9}$$

where $L(\theta) = \log \Pr(\zeta | \theta) = \log l(\theta | \zeta)$. A well known equivalent form of the fisher information matrix, which is defined based on the assumption of twice differentiability of the log-likelihood function, is:

$$F(\theta) \equiv -\mathbb{E} [H(\theta | \mathbf{z}) | \theta], \tag{10}$$

where

$$H(\theta | \zeta) \equiv \frac{\partial^2 L}{\partial \theta \partial \theta^T}$$

is the Hessian matrix. The latter form of the Fisher information matrix in (10) is often easier to compute than the former form given in (9). The analytical calculation of the Fisher information matrix is difficult here because of nonlinearity nature of the the epidemiological model. Thus, we use an approximation method based on a Monte Carlo sampling approach, which is introduced in (Spall 2005). The idea is to estimate the Hessian matrix of the log-likelihood function for a large number of samples and then average the negative of the Hessian matrices to obtain an estimate of the Fisher information matrix. Basically the same approach of simultaneous perturbation used in (4) can be used to estimate the Hessian matrix. A summary of the Monte Carlo sampling approach (from (Spall 2005)) is given here.

Let \mathbf{z}_k , $k = 1, \dots, K$, be the k th Monte Carlo generated sample from running the epidemic disease model using the estimated parameter, θ . To obtain \mathbf{z}_k , we need to run the simulation for the given θ . The pseudo measurement, \mathbf{z}_k , is a realization of the stochastic simulation based on the probability distributions for each node. The Monte Carlo based estimate of the information matrix is given by (Spall 2005):

$$\bar{F}_{J,k}(\theta) = \frac{k-1}{k} \bar{F}_{J,k-1}(\theta) - \frac{1}{kJ} \sum_{j=1}^J \hat{H}_{j|k}, \tag{11}$$

where $\hat{H}_{j|k}$ is the j th estimate of the Hessian matrix, $H(\theta)$ at the sample data \mathbf{z}_k . The inner averaging loop compute an estimate of the Hessian matrix at a sample data \mathbf{z}_k using J values of perturbations, $\Delta_1, \Delta_2, \dots, \Delta_J$. After executing (11) K times (for all samples \mathbf{z}_k), the $\bar{F}_{J,K}$ is used as an estimation of the Hessian matrix at θ . The details of the Monte Carlo based Hessian estimation can be found in (Spall 2005, Spall 2008).

4 PARALLEL SIMULTANEOUS PERTURBATION ALGORITHM: PSPO

In case of epidemiological model fitting, the objective function is obtained from (7). In order to compute this log likelihood function, we need to run the SSA simulation given θ to compute the probability term (8). Since SSA simulation is a stochastic process, the log likelihood function (the objective function) is

a noisy uncertain function. Although SPSA is an efficient optimization algorithm for high dimensional problems, it requires many iterations to converge, particularly in the case of high noise problems. Since all epidemic models are stochastic processes, the values of objective function are highly uncertain. In the case of highly uncertain problems, the computation of the gradient from noisy objective function require more runs to obtain an acceptable estimate. The idea, here, is using parallel computing of the gradient with different perturbation vectors. The multiple rounds of computation guarantees an error bound for each iteration given the noise variance presented in the objective function. The connection of the number of computation rounds and the error in estimated gradient is specified here.

The gradient vector at each iteration can be obtained by averaging over multiple evaluations of (3) for different values of perturbation. Let \mathbf{g}^i represents the directional gradient at a point θ in Δ_i direction, and \mathbf{g} is the gradient at this point. Then we know that:

$$\mathbf{g}^i = \frac{\langle \mathbf{g}(\theta), \Delta_i \rangle}{\|\Delta_i\|^2} \Delta_i.$$

Let $\hat{\mathbf{g}}^i$ is the estimated gradient using the one-sided gradient approximation (3) and perturbation vector Δ_i is a ± 1 vector. Then,

$$\hat{\mathbf{g}}^i \approx p\mathbf{g}^i = \langle \mathbf{g}(\theta), \Delta_i \rangle \Delta_i \approx \langle \hat{\mathbf{g}}(\theta), \Delta_i \rangle \Delta_i.$$

Then, we have

$$\hat{\mathbf{g}}^T \Delta_i \approx \frac{\delta f_i}{c}; \quad \delta f_i = L(\theta + c\Delta_i) - L(\theta),$$

and then,

$$\hat{\mathbf{g}}^T \Delta \approx \frac{1}{c} [\delta f_1 \quad \delta f_2 \quad \cdots \quad \delta f_M], \tag{12}$$

where,

$$\Delta = [\Delta_1 \quad \Delta_2 \quad \cdots \quad \Delta_M].$$

If $\{\Delta_i\}_{1 \leq i \leq M}$ span \mathbb{R}^p , then $\Delta\Delta^T$ is invertible. So, the least square estimation of $\hat{\mathbf{g}}$ is given by:

$$\hat{\mathbf{g}} \approx \frac{1}{c} (\Delta\Delta^T)^{-1} \Delta [\delta f_1 \quad \delta f_2 \quad \cdots \quad \delta f_M]^T. \tag{13}$$

On the other hand, if $M < p$, then (12) is an under-determined system which has non-unique solutions. Then, the solution of the minimum Euclidean norm, $\|\hat{\mathbf{g}}\|_2$, among all solutions is given by:

$$\hat{\mathbf{g}} \approx \frac{1}{c} \Delta (\Delta^T \Delta)^{-1} [\delta f_1 \quad \delta f_2 \quad \cdots \quad \delta f_M]^T. \tag{14}$$

The PSP algorithm for estimating the gradient at a given point is given in Algorithm 1. In order to improve the efficiency of this algorithm, we can compute $L(\theta)$ once, out of the loop. Doing this, we need $M + 1$ function evaluations for estimating the gradient.

One difference of the PSP algorithm compared with the conventional SPSA algorithm is on the selection of the perturbation vectors, Δ . We suggest a more efficient strategy to choose independent Δ vectors. First we sample Δ_0 from a Bernoulli ± 1 distribution. In round j , $1 \leq j \leq p$, we switch the sign of the j th element in Δ_0 to generate Δ_j . The same procedure repeats for the next p rounds of computations. In (Alaeddini and Klein 2017), it is proved that this strategy generates a set of perturbation vectors that spans \mathbb{R}^p .

We need to provide the number of parallel rounds of computation, M , as an input for this algorithm. Let $y = L(\theta)$ and $y^i = L(\theta + c\Delta_i)$ represent the noisy measurements at θ and $\theta + c\Delta_i$ respectively. Assume these measurements can be expressed as normal random variables as $y \sim \mathcal{N}(\mu, \sigma^2)$ and $y^i \sim \mathcal{N}(\mu_i, \sigma^2)$. It is assumed that the noise variance is constant over the whole state space, θ .

Inputs: given point θ , perturbation size c , and # of parallel rounds M ;

Randomly sample Δ_0 from $\{\pm 1\}$ binary distribution;

Initialize computation round counter $i := 1$;

repeat

$\bar{i} = i \bmod p$;
 $\Delta_i := (I - 2\mathbf{e}_{\bar{i}}\mathbf{e}_{\bar{i}}^T)\Delta_0$;
 $\delta f_i(\theta) := L(\theta + c\Delta_i) - L(\theta)$;
 $i := i + 1$;

until $i \leq M$;

$\Delta := [\Delta_1 \ \Delta_2 \ \dots \ \Delta_M]$;

if $M \geq p$ **then**

$\hat{\mathbf{g}}(\theta) := \frac{1}{c}(\Delta\Delta^T)^{-1}\Delta[\delta f_1 \ \delta f_2 \ \dots \ \delta f_M]^T$;

else

$\hat{\mathbf{g}}(\theta) := \frac{1}{c}\Delta(\Delta^T\Delta)^{-1}[\delta f_1 \ \delta f_2 \ \dots \ \delta f_M]^T$;

end if

Algorithm 1: PSP gradient.

Lemma 1 Given $y \sim \mathcal{N}(\mu, \sigma^2)$ and $y^i \sim \mathcal{N}(\mu_i, \sigma^2)$, if $M \geq p$, then the expected value of the gradient from Algorithm 1 is equal to the gradient \mathbf{g} .

Proof. Refer to (Alaeddini and Klein 2017). □

Theorem 1 If the rounds of computation, M , satisfies

$$M \geq \frac{\sigma^2 p}{c^2 \varepsilon^2}, \tag{15}$$

then the error in estimated gradient from Algorithm 1 is bounded by $\mathbb{E}[\|\hat{\mathbf{g}} - \mathbf{g}\|] \leq \varepsilon$.

Proof. Refer to (Alaeddini and Klein 2017). □

It is worthy to note that from the results of Theorem 1 one can show that if $M \rightarrow \infty$, then the error, $\varepsilon \rightarrow 0$, which is equivalent to $\mathbb{E}[\|\hat{\mathbf{g}} - \mathbf{g}\|] = 0$. Algorithm 2 presents a step-by-step summary of the PSPO algorithm, which is a conjugate gradient type algorithm. Details can be found in (Alaeddini and Klein 2017).

Running the epidemiological simulations, particularly the computation of likelihood function for a complex model, is usually an expensive task. Therefore, it is more desirable to have a robust solution, that we do not require to re-run the whole simulation with any change in the model. On the other hand, we would like to have a region of all possible solutions. Then we would be able to sample from the solution region and only keep the parameters that often produce good runs for the given observation. In order to estimate the solution region, we use the Monte Carlo based approach, introduced in (Spall 2005), and described earlier in this paper. At this point, we have all required tools to find the best predictive model from observed clinical data. Algorithm 3 presents a step-by-step summary of the proposed approach.

5 SIMULATION

To demonstrate the main contribution of this paper, we implemented the methodology on Haggeloch data set. This data set concerns a measles outbreak in a small town in Germany in 1861, which contains 188 infected individuals. This data set is very popular in literature because of its completeness and depth of data. Fig. 1 gives the observed clinical data. The data are obtained from (Meyer et al. 2014).

Initialize $i = 0$, $k = 0$ and initial guess θ_0 , Initialize update direction $\mathbf{d} = \text{null}$;
 Run PSP to compute $\hat{\mathbf{g}}_0(\theta_0)$ with $M = 1$;
 $\mathbf{r} := -\hat{\mathbf{g}}_0(\theta_0)$, $\mathbf{d} := \mathbf{r}$, $\mathbf{g}_{new} := \hat{\mathbf{g}}_0(\theta_0)$;
repeat
 Set # of parallel rounds M using error tolerance for this iteration;
 Run PSP to compute $\hat{\mathbf{g}}_k(\theta_k)$ using computed M ;
 $\mathbf{r} := -\hat{\mathbf{g}}_k(\theta_k)$, $\mathbf{g}_{old} := \mathbf{g}_{new}$, $\mathbf{g}_{new} := \hat{\mathbf{g}}_k(\theta_k)$;
 $\beta = \frac{\mathbf{g}_{new}^T(\mathbf{g}_{new} - \mathbf{g}_{old})}{\mathbf{g}_{old}^T \mathbf{g}_{old}}$;
 $\tilde{\mathbf{d}} := \frac{\mathbf{d}}{\|\mathbf{d}\|} + \varepsilon \mathbb{1}_{\mathbf{d}=0}(\mathbf{d})$;
 $\theta_k^+ := \theta_k + \tilde{\mathbf{d}}$, $\theta_k^- := \theta_k - \tilde{\mathbf{d}}$;
 Run PSP to compute $\hat{\mathbf{g}}(\theta_k^+)$, and $\hat{\mathbf{g}}(\theta_k^-)$, and compute $\delta G_k = \hat{\mathbf{g}}(\theta_k^+) - \hat{\mathbf{g}}(\theta_k^-)$;
 $\hat{H}_k := \frac{1}{2} \left[\frac{\delta G_k}{2} \tilde{\mathbf{d}}^{-1} + \left(\frac{\delta G_k}{2} \tilde{\mathbf{d}}^{-1} \right)^T \right]$;
 $\alpha := -\frac{\hat{\mathbf{g}}_k^T \mathbf{d}}{\mathbf{d}^T \hat{H}_k \mathbf{d}}$;
 Update $\theta_{k+1} := \theta_k + \alpha \mathbf{d}$, and update direction $\mathbf{d} := \mathbf{r} + \beta \mathbf{d}$;
 $i := i + 1$;
 if $i = p$ or $\mathbf{r}^T \mathbf{d} \leq 0$ **then**
 $\mathbf{d} := \mathbf{r}$;
 $i := 0$;
 end if
 $k := k + 1$;
until stopping criteria satisfied;

Algorithm 2: PSPO Algorithm.

Inputs: clinical data, $\zeta_1, \zeta_2, \dots, \zeta_n$;
 Initialize with given guess θ_0 ;
 Run PSPO algorithm until the stopping criteria satisfied;
 $\theta_{est} := \theta_k$ obtained from PSPO;
 Compute $\bar{F}(\theta_{est})$ using (11);
 $\Sigma_{\theta_{est}} = \bar{F}^{-1}(\theta_{est})$;
 Sample from θ_{est} and $\Sigma_{\theta_{est}}$ to find a desirable set of parameters;

Algorithm 3: Epidemic model calibration algorithm.

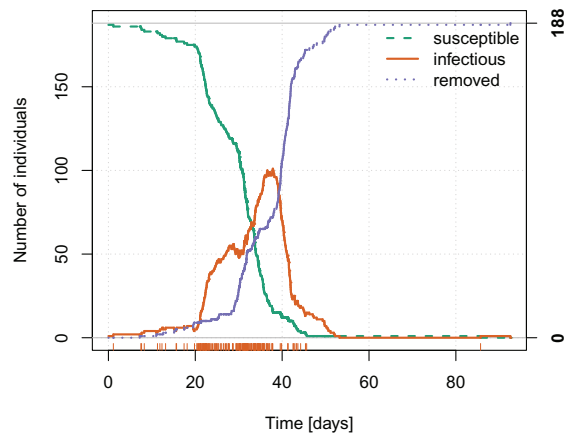


Figure 1: Measles outbreak clinical data.

Some types of diseases with permanent immunity, such as measles, mumps and rubella, can be described as Susceptible-Infected-Recovered (SIR) model. In this model each individual can only exist in one of the discrete states such as susceptible (S), infected (I) or permanently recovered (R). We have two transitions in this case. An infected person can infect others with an infection rate, β , and is cured with curing rate, δ . Therefore, the parameters of our model are $\theta = (\beta, \delta)$. Our objective, here, is finding a good model for the given epidemic data, i.e. infection rate and curing rate, to investigate the properties of the disease spread. These properties will allow the researchers to learn about the diseases, and thereby enabling them to test competing theories about transmission of disease and to devise better containment strategies.

We compare the efficiency of the PSPO algorithm with the conventional SPSA algorithm. To compare these two algorithms, we run the optimization algorithms 100 times for each one. Fig. 2a shows the number of iterations required to converge. As it can be seen in this figure, PSPO algorithm in average needs fewer iterations compared with the conventional SPSA algorithm. The maximum number of iterations for both algorithms set to 30. Note that the number of iterations until convergence set to 30 if the algorithm does not converge after 30 iterations. In order to investigate the performance of the PSPO algorithm as the number of parallel rounds increases, we run the PSPO for this example with different number of parallel computation rounds. Fig. 2b shows the number of iterations for convergence as M grows.

The comparison between SPSA and PSPO algorithm demonstrates the superiority of the PSPO algorithm in the number of iterations for convergence. Note that, comparing with SPSA, the PSPO algorithm with multiple parallel computations requires additional processing time per iteration. In order to prevent unnecessary long processing time per iteration, which results in long total processing time, the number of parallel computations need to be carefully computed using the level of the noise presented in the problem. In some cases when we are dealing with a high SNR (signal to noise ratio) problem, a single round of computation might be enough for convergence, but in some other applications, like most of the epidemiological applications, the data are highly noisy, and SPSA requires too many iterations to converge.

6 CONCLUSIONS

In many stochastic optimization problems that noise presents, obtaining an analytical solution is hardly possible. In this paper, we described an algorithm for optimal parameter estimation in discretely observed stochastic epidemic model. We presented an efficient algorithm which explores the parameters space to find a set of parameters which gives highest conditional probability of the given noisy measurements. The log likelihood function, which was used to evaluate the cost corresponding to different choices of parameters, was estimated by running the Gillespie stochastic simulation algorithm and then using the beta binomial

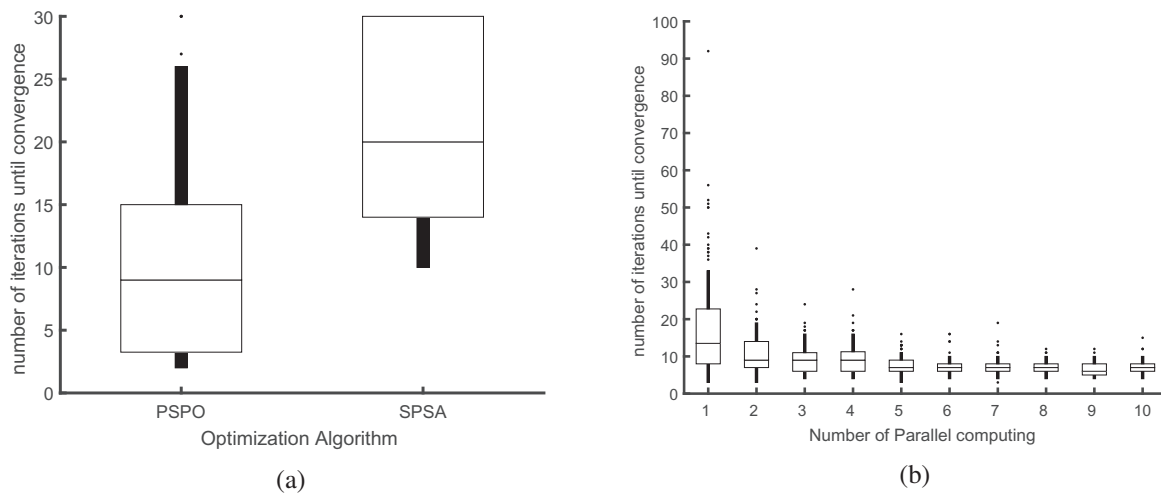


Figure 2: (a) Comparison of the number of iterations for convergence for conventional SPSA and PSPO. (b) Number of iterations vs. number of parallel computing rounds, M , in PSPO.

probabilistic model. We developed a parallel stochastic perturbation optimization algorithm and found the minimum number of parallel rounds of computation in order to have a bounded error in the estimated gradient. Furthermore, using a Monte Carlo re-sampling method, we computed the fisher information matrix, which gives the confidence region of the obtained solution.

ACKNOWLEDGMENTS

The authors thank Bill and Melinda Gates for their active support and their sponsorship through the Global Good Fund. The authors would also like to thank Philip Eckhoff for his helpful comments on the paper.

REFERENCES

- Alaeddini, A., and D. J. Klein. 2017. “PSPO: parallel simultaneous perturbation optimization”. *arXiv preprint arXiv:1704.00223*.
- Alaeddini, A., and K. A. Morgansen. 2016. “Optimal disease outbreak detection in a community using network observability”. In *American Control Conference (ACC)*, 7352–7357. IEEE.
- Andrianakis, I., I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White. 2015. “Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda”. *PLoS Comput Biol* 11 (1): e1003968.
- Bayati, B. S., and P. A. Eckhoff. 2012. “Influence of high-order nonlinear fluctuations in the multivariate susceptible-infectious-recovered master equation”. *Physical Review E* 86 (6): 062103.
- Bershteyn, A., D. J. Klein, and P. A. Eckhoff. 2016. “Age-targeted HIV treatment and primary prevention as a ring fence to efficiently interrupt the age patterns of transmission in generalized epidemic settings in South Africa”. *International health* 8 (4): 277–285.
- Bhatnagar, S., M. C. Fu, S. I. Marcus, I. Wang et al. 2003. “Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences”. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 13 (2): 180–209.
- Byrd, R. H., G. M. Chin, W. Neveitt, and J. Nocedal. 2011. “On the use of stochastic hessian information in optimization methods for machine learning”. *SIAM Journal on Optimization* 21 (3): 977–995.
- Eaton, J. W., N. Bacaër, A. Bershteyn, V. Cambiano, A. Cori, R. E. Dorrington, C. Fraser, C. Gopalappa, J. A. Hontelez, L. F. Johnson et al. 2015. “Assessment of epidemic projections using recent HIV survey

- data in South Africa: a validation analysis of ten mathematical models of HIV epidemiology in the antiretroviral therapy era". *The Lancet Global Health* 3 (10): e598–e608.
- Eckhoff, P. 2013. "Mathematical models of within-host and transmission dynamics to determine effects of malaria interventions in a variety of transmission settings". *The American journal of tropical medicine and hygiene* 88 (5): 817–827.
- Eckhoff, P. A., E. A. Wenger, H. C. J. Godfray, and A. Burt. 2016. "Impact of mosquito gene drive on malaria elimination in a computational model with explicit spatial and temporal dynamics". *Proceedings of the National Academy of Sciences*:201611064.
- Fröberg, C.-E. 1969. *Introduction to numerical analysis*. Addison-Wesley Publishing Company Reading Massachusetts.
- Fu, M. C. 1994. "Optimization via simulation: A review". *Annals of operations research* 53 (1): 199–247.
- Gerardin, J., C. A. Bever, B. Hamainza, J. M. Miller, P. A. Eckhoff, and E. A. Wenger. 2016. "Optimal population-level infection detection strategies for malaria control and elimination in a spatial model of malaria transmission". *PLoS Comput Biol* 12 (1): e1004707.
- Gillespie, D. T. 2001. "Approximate accelerated stochastic simulation of chemically reacting systems". *The Journal of Chemical Physics* 115 (4): 1716–1733.
- Grassly, N. C., C. Fraser, J. Wenger, J. M. Deshpande, R. W. Sutter, D. L. Heymann, and R. B. Aylward. 2006. "New strategies for the elimination of polio from India". *Science* 314 (5802): 1150–1153.
- Hong, L. J., and B. L. Nelson. 2009. "A brief introduction to optimization via simulation". In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 75–85. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Kiefer, J., J. Wolfowitz et al. 1952. "Stochastic estimation of the maximum of a regression function". *The Annals of Mathematical Statistics* 23 (3): 462–466.
- Klein, D. J., M. Baym, and P. Eckhoff. 2014. "The separatrix algorithm for synthesis and analysis of stochastic simulations with applications in disease modeling". *PloS one* 9 (7): e103467.
- Kocsis, L., and C. Szepesvári. 2006. "Universal parameter optimisation in games based on SPSA". *Machine learning* 63 (3): 249–286.
- Ma, J., Y. M. Nie, and H. M. Zhang. 2013. "Solving the integrated corridor control problem using simultaneous perturbation stochastic approximation". In *Advances in Dynamic Network Modeling in Complex Transportation Systems*, 89–113. Springer.
- Marshall, J. M., M. Touré, A. L. Ouédraogo, M. Ndhlovu, S. S. Kiware, A. Rezai, E. Nkhama, J. T. Griffin, T. D. Hollingsworth, S. Doumbia et al. 2016. "Key traveller groups of relevance to spatial malaria transmission: a survey of movement patterns in four sub-Saharan African countries". *Malaria journal* 15 (1): 200.
- McCarthy, K. A., G. Chabot-Couture, and F. Shuaib. 2016. "A spatial model of Wild Poliovirus Type 1 in Kano State, Nigeria: calibration and assessment of elimination probability". *BMC Infectious Diseases* 16 (1): 521.
- Meyer, S., L. Held, and M. Höhle. 2014. "Spatio-temporal analysis of epidemic phenomena using the R package surveillance". *arXiv preprint arXiv:1411.0416*.
- Miller, J. C., A. C. Slim, and E. M. Volz. 2012. "Edge-based compartmental modelling for infectious disease spread". *Journal of the Royal Society Interface* 9 (70): 890–906.
- Polyak, B. T., and A. B. Juditsky. 1992. "Acceleration of stochastic approximation by averaging". *SIAM Journal on Control and Optimization* 30 (4): 838–855.
- Preciado, V. M., M. Zargham, C. Enyioha, A. Jadbabaie, and G. Pappas. 2013. "Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks". In *IEEE 52nd Annual Conference on Decision and Control*, 7486–7491.
- Rao, C. R. 1992. "Information and the accuracy attainable in the estimation of statistical parameters". In *Breakthroughs in statistics*, 235–247. Springer.

- Robbins, H., and S. Monro. 1951. "A stochastic approximation method". *The annals of mathematical statistics*:400–407.
- Roh, M. K., and P. Eckhoff. 2014. "Stochastic parameter search for events". *BMC systems biology* 8 (1): 126.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1985. "Learning internal representations by error propagation". Technical report, DTIC Document.
- Scharf, L. L. 1991. *Statistical signal processing*, Volume 98. Addison-Wesley Reading, MA.
- Spall, J. C. 1992. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". *IEEE transactions on automatic control* 37 (3): 332–341.
- Spall, J. C. 2000. "Adaptive stochastic approximation by the simultaneous perturbation method". *IEEE transactions on automatic control* 45 (10): 1839–1853.
- Spall, J. C. 2005. "Monte carlo computation of the fisher information matrix in nonstandard settings". *Journal of Computational and Graphical Statistics* 14 (4): 889–909.
- Spall, J. C. 2008. "Improved methods for monte carlo estimation of the fisher information matrix". In *American Control Conference*, 2395–2400. IEEE.
- Spall, J. C. 2009. "Feedback and weighting mechanisms for improving Jacobian estimates in the adaptive simultaneous perturbation algorithm". *IEEE Transactions on Automatic Control* 54 (6): 1216–1229.
- Spall, J. C. 2012. "Stochastic optimization". In *Handbook of computational statistics*, 173–201. Springer.
- Swisher, J. R., P. D. Hyden, S. H. Jacobson, and L. W. Schruben. 2000. "A survey of simulation optimization techniques and procedures". In *Proceedings of Winter Simulation Conference*, Volume 1, 119–128. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Tsilifis, P., R. G. Ghanem, and P. Hajali. 2017. "Efficient bayesian experimentation using an expected information gain lower bound". *SIAM/ASA Journal on Uncertainty Quantification* 5 (1): 30–62.
- Wagner, B. G., M. R. Behrend, D. J. Klein, A. M. Uphill-Brown, P. A. Eckhoff, and H. Hu. 2014. "Quantifying the impact of expanded age group campaigns for polio eradication". *PloS one* 9 (12): e113538.
- Wright, S., and J. Nocedal. 1999. *Numerical optimization*, Volume 2. Springer New York.

AUTHOR BIOGRAPHIES

ATIYE ALAEDDINI is a Postdoctoral Research Scientist at the Institute for Disease Modeling, Applied Math team. She holds a Ph.D. in Aerospace engineering from University of Washington and a M.Sc. in Computer engineering from University of California, Santa Cruz. Her research interests include stochastic optimization, control theory, online optimization algorithms, networked systems and their applications in epidemiology. As a member of IDMs research team, she is developing a statistical model and optimization algorithm for model calibration and parameter space exploration. Her e-mail address is aalaeddini@idmod.org.

DANIEL KLEIN is Sr. Research Manager at the Institute of Disease Modeling. He holds both Ph.D. and M.Sc. in Aerospace engineering from University of Washington. As the head of the Applied Math team at IDMs research team, his time is split between developing the HIV model and building research tools. His HIV work focuses on building the contact network using ideas from feedback control to guide the dynamical relationship formation process. He is exploring structural assumptions in HIV modeling, HIV model calibration, the impact of interventions, and parameter sensitivity. His email address is dklein@idmod.org.